DOCUMENT RESUME

ED 333 020 TM 016 495

AUTHOR Hecht, Jeffrey B.
TITLE Least-Squares Linear Regression and Schrodinger's Cat: Perspectives on the Analysis of Regression Residuals.
PUB DATE Apr 91
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; Evaluation Methods; *Least Squares Statistics; *Regression (Statistics)
IDENTIFIERS DrawReg Computer Program; *Outliers; *Residuals (Statistics); Schrodingers Cat

ABSTRACT

The analysis of regression residuals and detection of outliers are discussed, with emphasis on determining how deviant an individual data point must be to be considered an outlier and the impact that multiple suspected outlier data points have on the process of outlier determination and treatment. Only bivariate (one dependent and one independent) models were investigated. A sufficient number of data points (240 pairs) was used to avoid interpretation problems associated with small data sets. The original data set was successively manipulated to include additional data points with increasingly larger degrees of extremeness, and potential outlier data points were added. To meet the demands of the data set modification, a computer program, DrawReg, was written in QuickBASIC. Outlier points could be viewed as belonging to one of five broad categories, and it was apparent that no single residual statistic could adequately account for all types. The use of multiple outlier detection techniques is recommended as part of the least-squares modeling process. The importance of the argument is illustrated through the discussion of the Schrodinger's Cat situation derived from quantum physics. Twelve tables and six figures illustrate the discussion. An appendix describes the DrawReg program. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED333020

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JEFFREY B. HECHT
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Least-Squares Linear Regression and Schrodinger's Cat: Perspectives on the Analysis of Regression Residuals

by

Jeffrey B. Hecht
Department of Educational Administration and Foundations
College of Education
Illinois State University

BEST COPY AVAILABLE

This paper was presented at the Annual Meeting
of the
American Educational Research Association
April 3-7, 1991
Chicago, Illinois

TM016495

The Way it Began

It all started very simply, as interesting problems usually do.

Not too long ago I was teaching an Introductory Statistical Methods course, having reached that part of the plan which dealt with the creation of bivariate least-squares regression models. The students had just completed the determination of a model for the practice data set being used, and I was discussing ways in which such models might be interpreted. I had already addressed ascertaining the adequacy of a linear model (validating the key assumptions), and was working up towards an assessment of the fit of the model to the data set.

"Before you can begin any interpretation", I began, "it is important to determine whether any data points are exerting an undue influence on your model. If such data points exist they might bias the parameters, goodness of fit, and interpretation of your linear model."

I then undertook a discussion of the calculation of residuals, and the manipulation of residuals into different statistics (standardized residuals, studentized residuals, jackknifed residuals, leverage, and Cook's Distance). The data set in question had a single data point with a huge relative residual, making it easy to identify that point as the potential outlier under scrutiny. The usual discussion surrounding what to do with that point ensued.

Then it happened.

One of the students asked, "How far 'out' does a data point have to be before you should consider it an outlier?"

Then another, "What do you do if there's more than one?"

Introduction

Regression analysis represents one statistical technique widely used for modeling real-world phenomenon. Regression techniques are frequently employed in exploratory modeling, where a search is made for independent variables that contribute significant explanatory power towards the prediction of a dependent variable. They are also used in confirmatory analyses, where gathered data is fitted against a theoretically-derived model that (hopefully) captures the relationship of the various independent variables to the dependent (Kleinbaum, Kupper, and Muller, 1988, pp. 36-40).

In both cases there is an understanding that even the best of models will not fit the data perfectly (Hays, 1988, p. 545). Error, or the degree of mismatch between the values the model predicts and those actually observed, are to be expected in any real-world process. Too much error, however, may represent flaws in either the definition of data elements or in their collection strategy. Likewise the model itself may be flawed, either by representing relationships which are not reflected in the data or by violating one (or more) of the key assumptions necessary under the model's distribution theory (Harnett, 1982, pp. 487-493).

Existing methods present a variety of ways for identifying the overall "goodness of fit" of any given model. F ratios, R^2, adjusted R^2 and various other G.F.I. indices serve as summary indicators of how well a particular model reproduces the data (Norusis, 1990, pp. B-74-B-76). When the fit of a particular model, especially the more common least-squares approach that is designed to minimize error under assumed linearity, is judged as less-than-adequate, additional steps are usually employed to seek an improved fit.

In this vein the experimenter usually begins with various tests designed to ascertain the appropriateness of the application of a linear least-squares to the data set in question. These tests include an assessment of linearity, equality of variance (or homoscedasticity), normal distribution, and independence of the error terms from each other and the independent variable(s). Moderate violations of these assumptions can result in a misfit model and uncertain predictable power.

Even when these assumptions are satisfactorily met, however, problems can arise that significantly impact the interpretation of the proposed linear model. Outliers, those individual data points whose influence on the parameters of the model is unusually large, may exist. Such data points, if they exist in a real data set, may exert an undue influence on either the estimated values of the parameters of the model, the interpretation of those parameters, or both (Stevens, 1986, p. 86).

Typical methods used in the identification of outliers include: the unstandardized residual, standardized residual, the Student's t residual, jackknife residual (and, by expansion, bootstrapped residuals), and the calculation of leverage (including Cook's Distance). Plotting of the residuals, especially in reference to one (or more) of the other variables or the passage of time, is also a popular way to ascertain which data points are more deviant than others. Regardless of the method, however, the aim of such exercises remains the same; namely, to identify that point(s) which "is a peculiarity ... a data point which is not at all typical of the rest of the data" (Draper & Smith, 1981, p. 152).

This paper seeks to clarify the discussion of residual analysis and outlier detection by asking several questions central to the topic. First, how deviant must an individual data point be, from either the centroid of its set or some predicted line, before it is declared to be an outlier? The process of such identification, at least as outlined in the major texts on the subject, appears somewhat unclear. Second, and as important, is what impact multiple suspected outlier data points have on the process of outlier determination and treatment? There is the suggestion in current texts that the procedures for multiple outliers ought be similar, if not identical to, those for a single suspected case. An argument can be made, however, for a somewhat different approach.

Ground Rules for this Investigation

Certain ground rules were adopted from the outset to maintain the central points of this discussion without getting unnecessarily distracted by peripheral issues. First, only bivariate (one dependent and one independent) models were investigated. Though outliers do exist in multi-variable models such a consideration would, it was felt, only complicate the discussion unnecessarily. Likewise, all data sets considered conformed to the standard "HEIL GAUSS" (Homoscedasticity, Existence, Independence, Linearity, and Gaussian distribution) assumptions (Kleinbaum, Kupper, & Muller, 1988, p. 49). While some deviation had to be introduced, in the form of the outlier data point(s) being considered, these assumptions were still significantly satisfied in all data sets.

More importantly, it was assumed that all data points had already been verified as plausible within the context of their data set. In this regard the admonitions of Draper and Smith (1981) are quite sound: "As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors in recording the observations or in setting up the apparatus" (p. 153). The data points in each data set were therefore verified as being within the ranges realistically possible for such data. A negative value for a person's height, for example, would clearly be incorrect and would be removed from the data set before any analyses were performed.

Method

The basic approach used in this research was to start with a bivariate data set exhibiting a moderate degree of variability in both the X and Y measures. A sufficient number of data points (240 pairs) were used so problems of small data set interpretation could be sidestepped. The data points correlated moderately (r=.5165) and, at the outset, contained no points that would be construed as extreme outliers.

This original data set was then successively manipulated so as to include additional data points which exhibited increasingly larger degrees of extremeness. Data pairs were included that varied from the X centroid only, the Y only, and both the X and Y common center. At each successive data set basic statistics were calculated and a plot of the regression drawn.

Eventually potential outlier data points were added to the original data set several at a time, with varying degrees of "extremeness" among the data points added. Different methods were applied for both identifying the potential outliers in the data set, and then for reducing the data set to eliminate their influence. The difficulty in implementation, and success of the methods at achieving both point identification and less-influenced subset equation, were noted at each step. Over 500 different data sets were considered, though only a few are presented as examples in this paper.

The original research began by using SPSS/PC+ (Norusis, 1990) as the computer program to estimate the least squares model, compute the residual diagnostic statistics, and draw the plots. As the research progressed, however, it became clear that this package was not well suited to producing the kind of immediate feedback and quality screen graphics that were desired as the data set was successively modified. An investigation of other commercially available packages was likewise unsuccessful at providing a sufficient instrument.

For this reason a new computer program was written to satisfy the needs of fast data set case-by-case manipulation (addition & deletion), residual statistical computation, and graphical presentation. "DrawReg" was written in QuickBASIC (Microsoft Corporation, 1989). It runs on any PC/XT, AT, or 386-based computer having some graphics display capability (CGA, EGA, VGA, or HGC). 640Kb of memory is needed and a math coprocessor is recommended to improve the program's running speed (though it is not required for basic operation). Results from DrawReg were validated against those obtained from SPSS/PC+ (version 4.0) on similar data sets. The graphics and tables included in the remainder of this report were produced through DrawReg on an PC/AT computer with a VGA monitor. Additional information on this program is available in Appendix A.

Results

The Original Data Set

The regression statistics calculated for the initial data set are presented below in Table 1.

	First Data Set
Cases	240
X Values: Mean	68.44917
Sdev	2.58672
Y Values: Mean	171.97500
Sdev	22.34794
B0 (Intercept)	-133.51172
B1 (Slope)	4.46297
Serr of Y|X	19.17535
r (Correlation)	0.51658
R2 (Coeff of D)	0.26685

Table 1: Initial Data Set Regression Statistics

Residual analysis was performed using four measures: standardized residuals, studentized residual, leverage, and Cook's distance. Three data points were identified as potentially too far leveraged, twelve had values beyond the 1.96z typically employed, and no pairs were found to exceed the critical value for Cook's statistic (see Table 2). Examination of the plot, however, finds this data set to be relatively well-ordered (Figure 1).

Potential X outlier cases only (Leverage order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
128		76.000	.0398	222.000	205.674	0.851	0.869	0.016
21		60.900	.0398	149.000	138.283	0.559	0.570	0.007
206		75.300	.0335	212.000	202.550	0.493	0.501	0.004

3 total cases displayed

Potential Y outlier cases only (Standardized Residual order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
140		71.500	.0100	235.000	185.591	2.577	2.590	0.034
154		72.200	.0130	141.000	188.715	-2.488	-2.505	0.041
194		69.300	.0046	223.000	175.772	2.463	2.469	0.014
153		68.200	.0042	125.000	170.863	-2.392	-2.397	0.012
118		68.100	.0042	215.000	170.417	2.325	2.330	0.012
28		69.000	.0044	219.000	174.433	2.324	2.329	0.012
176		71.800	.0112	228.000	186.930	2.142	2.154	0.026
65		67.500	.0047	127.000	167.739	-2.125	-2.130	0.011
91		70.800	.0076	143.000	182.467	-2.058	-2.066	0.016
226		72.300	.0134	150.000	189.161	-2.042	-2.056	0.029
150		70.000	.0057	140.000	178.896	-2.028	-2.034	0.012
201		67.500	.0047	206.000	167.739	1.995	2.000	0.010

12 total cases displayed

Potential X-Y outlier cases (Cooks Distance order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks

0 total cases displayed

Table 2: Initial Data Set Residual Statistics

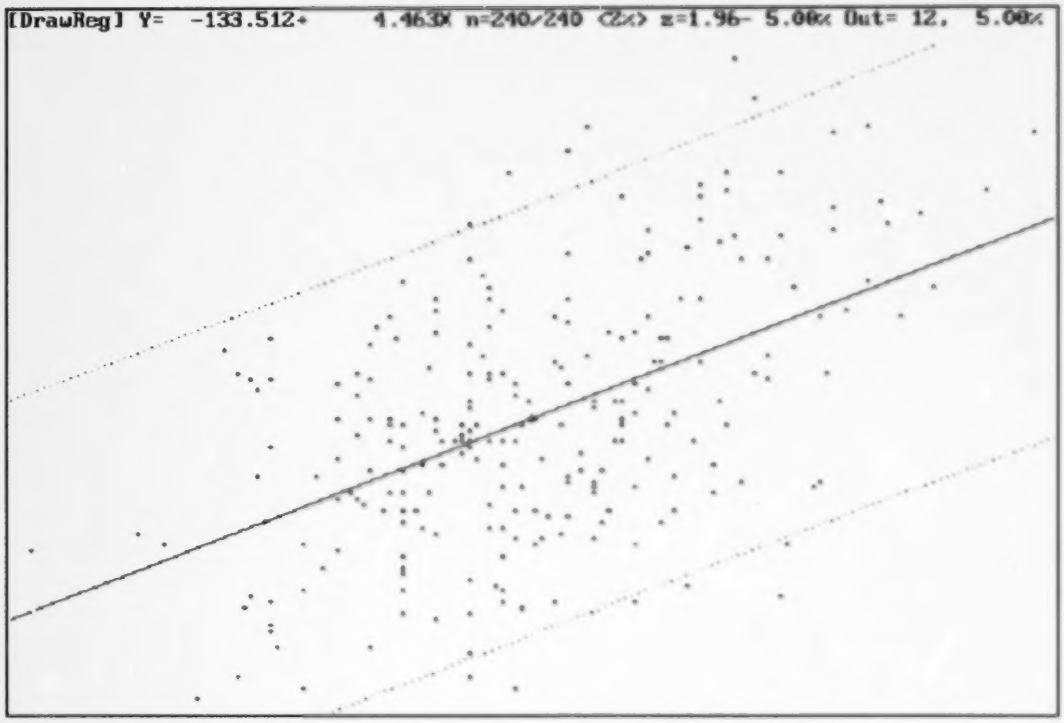

Figure 1: Initial Data Set (with Regression Line & 95% Pred Bands)

A Single Outlier

A single data point, of increasingly deviant X and Y values, was successively added into the initial data set. Regression and residual statistics were calculated at each iteration. After numerous runs were made, these statistics were compared to determine trends among the runs. Two important findings were made.

First, it is possible to characterize potential outlier cases as exerting primary influence over the dependent (Y) variable range, the independent (X) variable range, or over both. Outlier cases with a strong Y-variable influence (as identified through large values in both the standardized and studentized residuals) tended to most influence the parameters (and coefficient of determination) of the least-squares regression line. This was less evident when the outlier data point is located near the common centroids as when it is located further away. It is interesting to note that, while these points show a large standardized and studentized residual their leverage and Cook's distance values are well within the

accepted range. Figure 2 and Tables 3 and 4 give a typical example of a single data point with a large standardized (and studentized) residual but, due its location near the common X-Y center, a low leverage and Cook's Distance value.

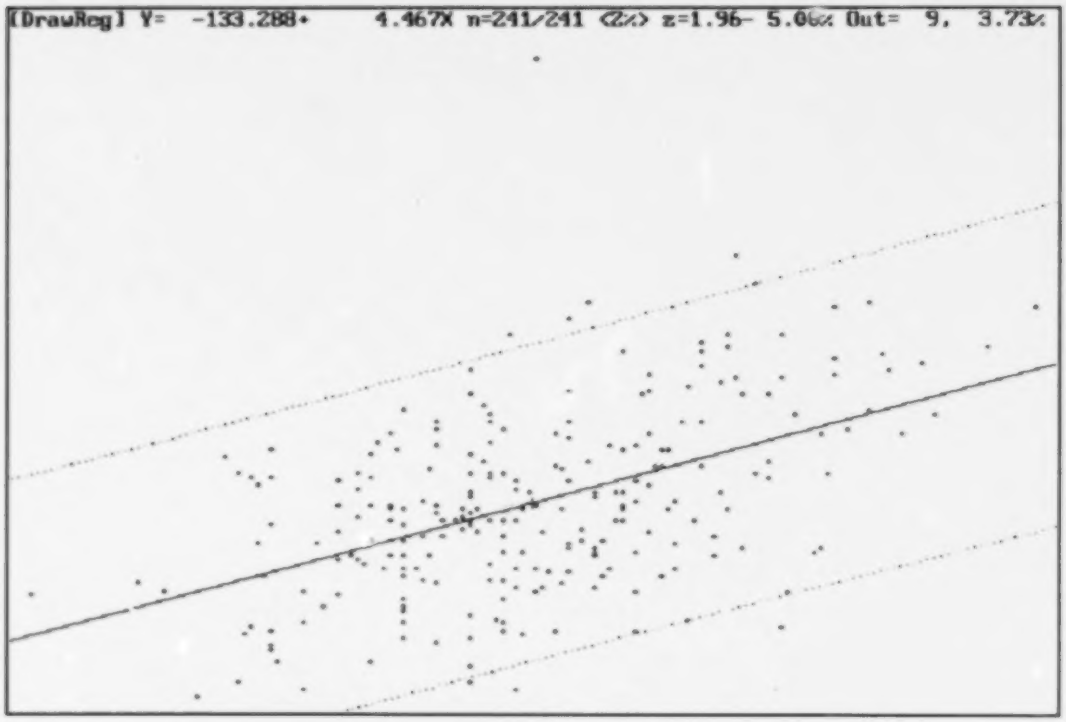

Figure 2: Data Set with One, Centered Y-Outlier

Potential Y outlier cases only (Standardized Residual order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
1		68.500	.0042	285.000	172.670	5.487	5.498	0.063

Table 3: Residual Statistics on One, Centered Y-Outlier

	This Data Set	First Data Set
Cases	241	240
X Values: Mean	68.44938	68.44917
Sdev	2.58133	2.58672
Y Values: Mean	172.44398	171.97500
Sdev	23.45967	22.34794
B0 (Intercept)	-133.28809	-133.51172
B1 (Slope)	4.46654	4.46297
Serr of Y|X	20.47366	19.17535
r (Correlation)	0.49147	0.51658
R2 (Coeff of D)	0.24154	0.26685

Table 4: Regression Statistics on One, Centered Y-Outlier

Data points showing a high leverage (deviant X value) tended to impact the regression statistics to a small degree. These points exhibited a large leverage, but within-range values for the other residual diagnostics. In many cases it was difficult to consider them an outlier from the regression standpoint, since they lay reasonably close to the predicted line. If anything they contributed more to the variability of the X data points than to the relationship between the variables or with the least-squares regression line.

These points, however, have the potential for exerting a great influence on the regression parameters. In the same way that a single, heavy youngster can balance on a teeter-totter with several smaller children on the other side, so can one well placed X-outlier influence an entire regression equation. What makes these points more difficult to locate is that they may show little (if any) deviation on the Y data set while meaningfully altering the least-squares estimation. Figure 3 and Tables 5 and 6 show one non-extreme example.

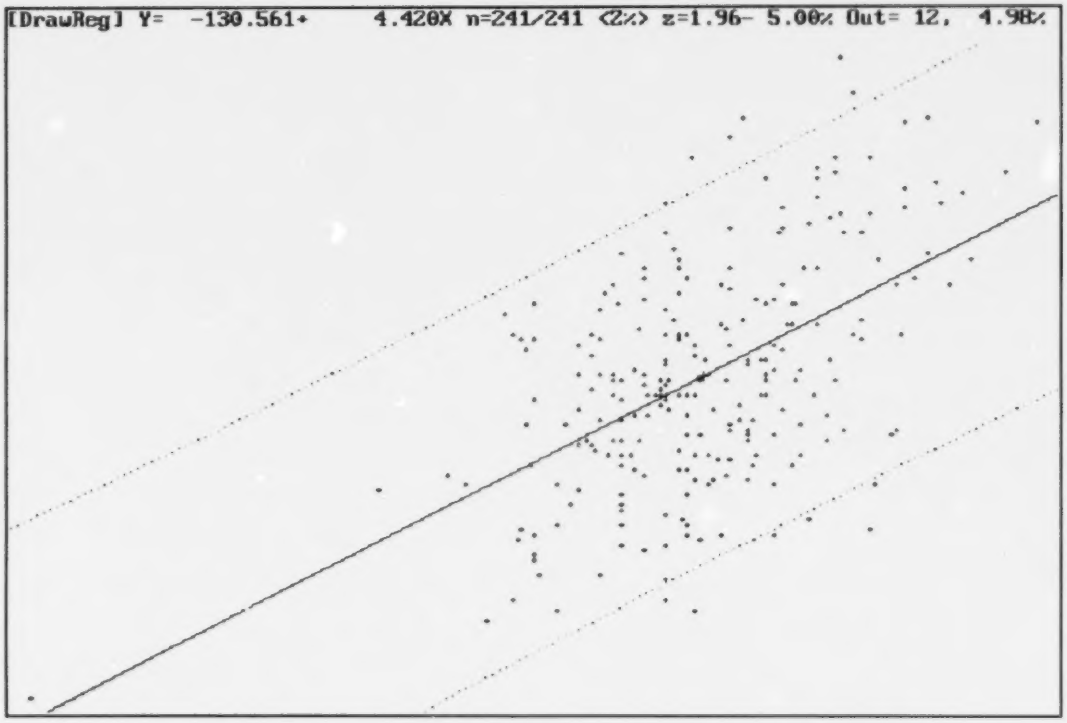

Figure 3: Data Set with One, Non-Extreme X-Outlier

Potential X outlier cases only (Leverage order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
		52.900	.1345	107.670	103.264	0.230	0.247	0.005

Table 5: Residual Statistics on One, Non-Extreme X-Outlier

	This Data Set	First Data Set
Cases	241	240
X Values: Mean	68.38465	68.44917
Sdev	2.76884	2.58672
Y Values: Mean	171.70817	171.97500
Sdev	22.68276	22.34794
B0 (Intercept)	-130.56097	-133.51172
B1 (Slope)	4.42013	4.46297
Serr of Y|X	19.13765	19.17535
r (Correlation)	0.53956	0.51658
R2 (Coeff of D)	0.29112	0.26685

Table 6: Regression Statistics on One, Non-Extreme X-Outlier

Most easily identified were those points having a relatively large deviation from both the X and Y centers. These points tended to show extreme values on all of the four residual measures being used, while also exerting a noticeable impact on the regression estimates.

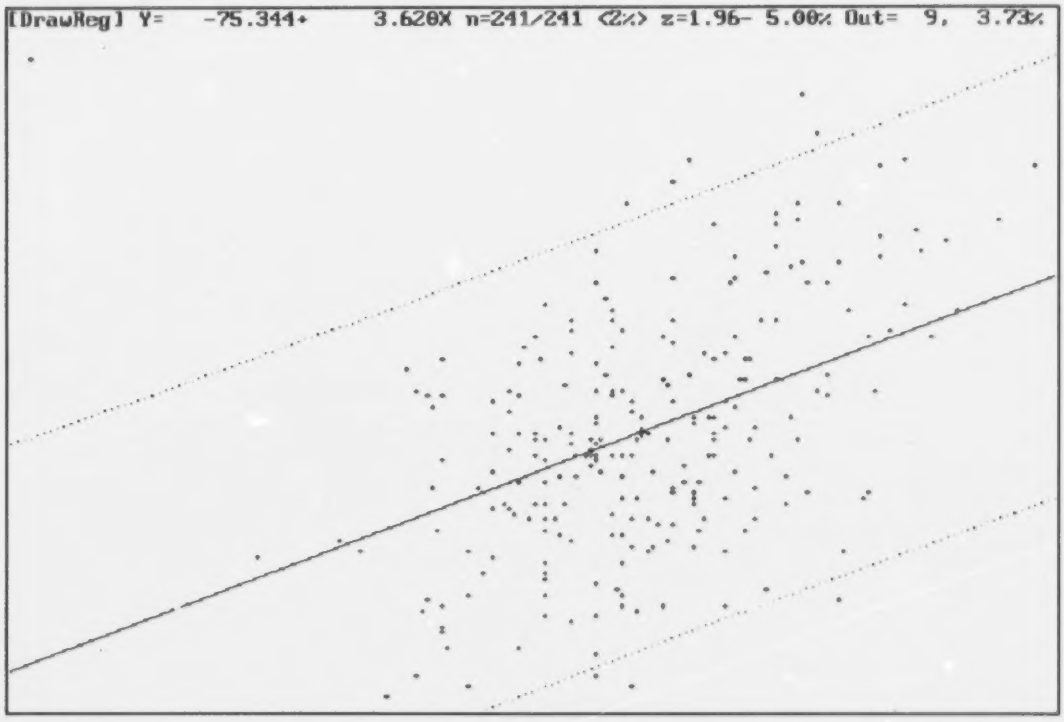

Figure 4: Data Set with One, Balanced Outlier

Potential outlier cases only

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
1		56.500	.0855	242.000	129.189	5.476	5.726	1.532

Table 7: Residual Analysis of One, Balanced Outlier

	This Data Set	First Data Set
Cases	241	240
X Values: Mean	68.39959	68.44917
Sdev	2.69364	2.58672
Y Values: Mean	172.26556	171.97500
Sdev	22.75293	22.34794
B0 (Intercept)	-75.34358	-133.51172
B1 (Slope)	3.62004	4.46297
Serr of Y|X	20.60049	19.17535
r (Correlation)	0.42856	0.51658
R2 (Coeff of D)	0.18367	0.26685

Table 8: Regression Statistics on One, Balanced Outlier

More Than One Outlier

The same basic pattern of outlier location was observed with multiple potential outlier cases. More than one point could be grouped together primarily off the Y centroid, the X centroid, or some balance between both. They also could be mixed between these three, in addition to appearing on both sides of the least-squares regression line. There are a limitless combination of possible potential outlier locations and distances from the centroids; however, certain types of combinations tended to stand out as especially interesting.

The first of these could be called a bipolar balanced outlier pair. In this simple case two points are located approximately equally distant from the common center but on opposing sides of the regression equation line. Thus, their combined influence cancels out any affect on the regression parameters other than to increase the variability of the Y points (and, therefore, decrease the correlation between X and Y) (Figure 5 and Tables 9 & 10).

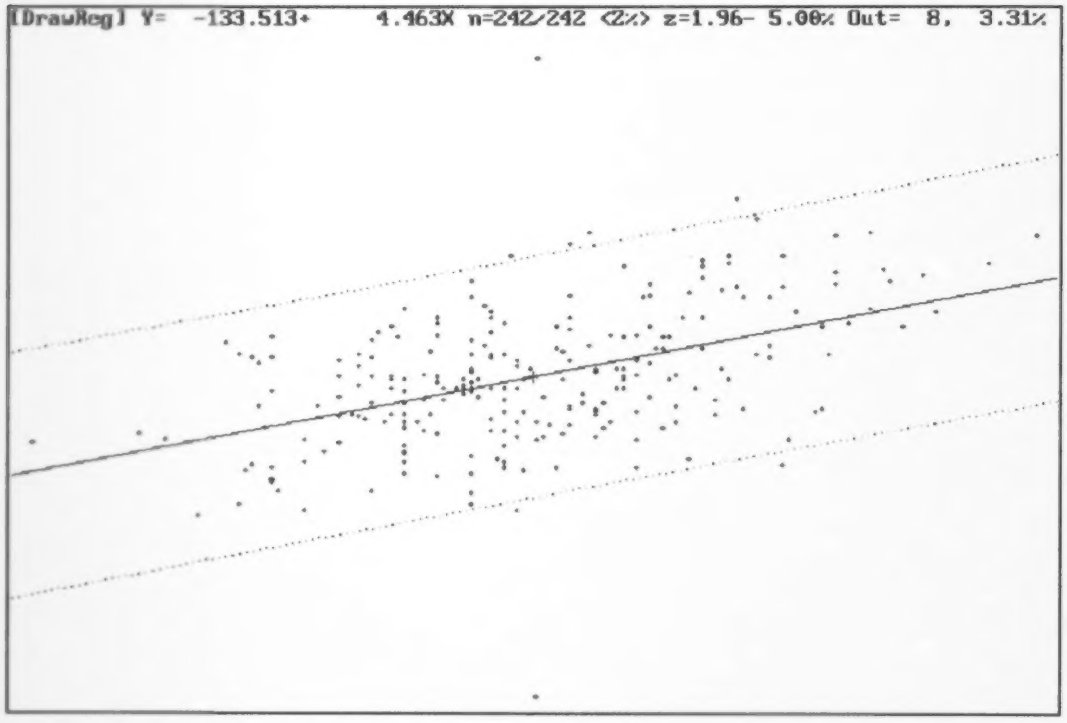

Figure 5: Data Set with One, Bipolar Balanced Outlier Pair

Potential Y outlier cases only (Standardized Residual order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
2		68.500	.0041	58.950	172.200	-5.218	-5.229	0.057
1		68.500	.0041	285.000	172.200	5.197	5.208	0.056

Table 9: Residual Statistics on One, Bipolar Balanced Outlier Pair

	Both Points In	One Point In	First Data Set
Cases	242	241	240
X Values: Mean	68.44959	68.44938	68.44917
Sdev	2.57597	2.58133	2.58672
Y Values: Mean	171.97500	172.44398	171.97500
Sdev	24.52141	23.45967	22.34794
B0 (Intercept)	-133.51262	-133.28809	-133.51172
B1 (Slope)	4.46296	4.46654	4.46297
Serr of Y|X	21.70451	20.47366	19.17535
r (Correlation)	0.46883	0.49147	0.51658
R2 (Coeff of D)	0.21980	0.24154	0.26685

Table 10: Regression Statistics on One, Bipolar Balanced Outlier Pair

The second interesting case also represents a canceling situation, but of a much different nature. Here, a single data point is well distanced from the common center on one side of the line but is balanced by several other data points much closer in yet on the opposite side (Figure 6 and Tables 11 & 12). Again, this combination of points produce little change in the regression parameters although, as will be seen in the discussion to follow, a challenge when examining traditional methodologies for addressing what to do with these outlier data points.

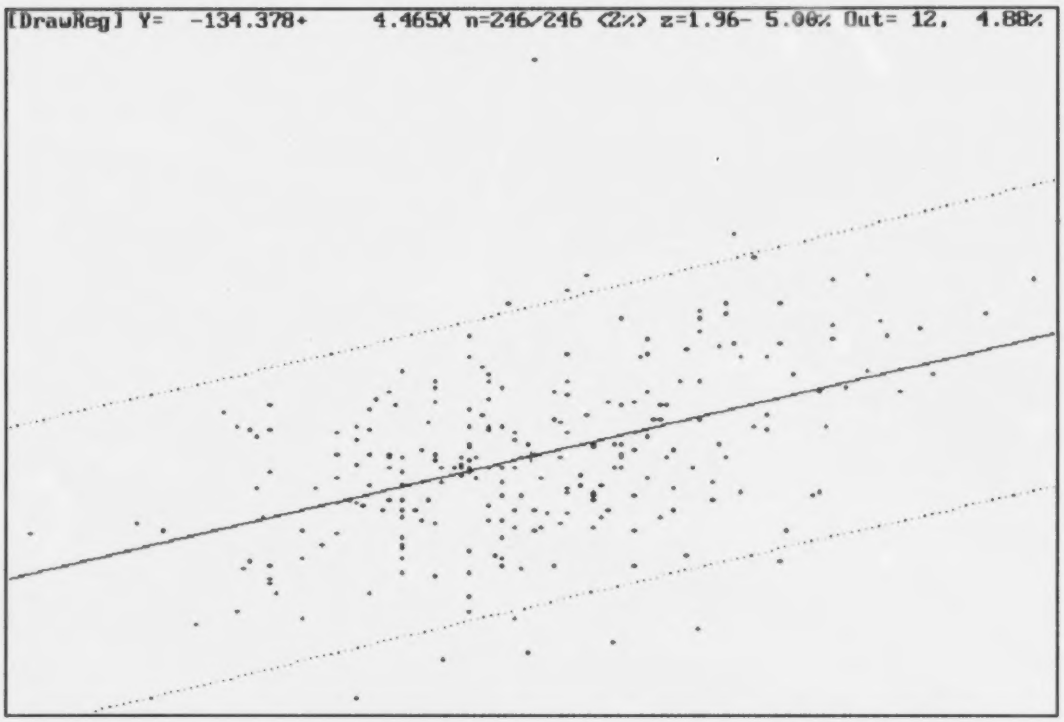

Figure 6: Data Set with Multiple, Balanced Outlier Points

All cases in the original data set (original order)

Obs	?	X-Value	(.033) Lever	Y-Value	Predicted	(1.96) Stand	(1.96) Stdnt	(1.00) Cooks
1		68.500	.0041	285.000	171.487	5.188	5.198	0.055
2		65.820	.0083	102.000	159.520	-2.629	-2.640	0.029
3		67.110	.0052	113.000	165.280	-2.389	-2.395	0.015
4		68.400	.0041	115.000	171.040	-2.561	-2.566	0.013
5		69.690	.0050	118.000	176.800	-2.687	-2.694	0.018
6		70.980	.0080	122.000	182.560	-2.768	-2.779	0.031

Table 11: Residual Statistics on Multiple, Balanced Outlier Points

	All Points In	One Point In	First Data Set
Cases	246	241	240
X Values: Mean	68.44837	68.44938	68.44917
Sdev	2.56812	2.58133	2.58672
Y Values: Mean	171.25610	172.44398	171.97500
Sdev	24.66451	23.45967	22.34794
B0 (Intercept)	-134.37850	-133.28809	-133.51172
B1 (Slope)	4.46518	4.46654	4.46297
Serr of Y|X	21.88144	20.47366	19.17535
r (Correlation)	0.46492	0.49147	0.51658
R2 (Coeff of D)	0.21615	0.24154	0.26685

Table 12: Regression Statistics on Multiple, Balanced Outlier Points

Further examples of combinations of groups of data points, at different geometric locations and distances from the common center, were also investigated. Regardless of the combinations examined they all appeared to exhibit the characteristics of the above mentioned five examples. It should be noted that the aggregation of too many unbalanced data points, while easily detected as outlier cases, also tend to violate the basic assumptions of the linear least-squares model. These violations are easily detected through other means before the regression model is even estimated.

Summary of the Results

This investigation has demonstrated that outlier data points, in the bivariate regression model, can be viewed as belonging to one of five broad categories. Single points may exhibit a major influence on the Y axis, major influence on the X axis, or a shared influence on both axes. To avoid violation of the least-squares assumptions multiple data points must "balance" across the estimated regression line: either in pairs of closely matched points or approximately equal geometric distances from the common X-Y center. While other combinations or outliers are possible, they should exhibit similar characteristics to these five categories.

Discussion

It is clear from this study that no single residual statistic will adequately account for all of the types of potential outliers. Standardized and studentized residual analysis proved most effective at identifying Y-axis outlier cases while the computation of leverage was helpful in the identification of X-axis outliers. Cook's distance was most sensitive to points deviant on both axes. None of the measures, however, was able to capture each distant point in all cases.

Such a realization argues strongly for the use of multiple outlier detection techniques as part of the least-squares modeling process. Employing techniques sensitive to each of the three first categories of outliers will reduce the risk of an extreme, influential data point going undetected. While this strategy improves outlier detection, it does not address the issue of what to do with those influential data points once they are found.

In this regard most writers argue that the starting point for any outlier analysis should begin at the plausibility of the data points in question. If the range of allowable values for a given data point is exceeded, such as a score of 110% for a test that can only range from 0% to 100%, an obvious error exists. The raw data representing these data need to be examined and, when coding errors exist, corrected. If they cannot be corrected (due to source coding problems) the researcher must decide whether to recode the data point to some extreme, yet allowable, value or to eliminate the point entirely.

Elimination of data is rarely a desirable option. Aside from the expense associated with collecting data most researchers want to insure that they are not unreasonably manipulating a data set just to make it conform with their assumptions. This is where real caution needs to be exercised.

Outlier data points that are not obvious coding errors have traditionally been examined according to some predefined statistical measure of adequacy (such as Z or t tables). Those points exceeding some researcher defined limit (typically, although perhaps mostly arbitrarily, set at $\alpha = .05$) have been declared to be outliers and, therefore, subject to recoding or elimination from the data set. Such a scheme ignores the basic assumptions upon which linear least-squares modeling is performed. Under the bivariate gaussian distribution a single data point has the possibility of existing *anywhere* in the within-range point "space". The contention that because a point is extremely located from the common center does not invalidate the necessary distributional assumptions. It is only when a meaningfully large grouping of such points exhibit the extreme tendencies do the assumptions come into question.

Schrodinger's Cat

The importance of this argument is perhaps best made by way of a tale attributed to the beginnings of quantum mechanics theory. At the beginning of this century Erwin Schrodinger was working through some of the complicated theoretical issues associated with this new conceptualization of physics. He imagined a situation which came to be known as "Schrodinger's Cat" (Rae, 1986, p. 59), a variation of which can be used to illuminate the discussion at hand.

Imagine a device is constructed which contains a single atom of some radioactive substance, a sensitive detector capable of determining when that atom has undergone radioactive decay, a gun (or some other lethal object), and a living cat. The detector is connected to the gun in such a way that when the atom is measured as having undergone radioactive decay the gun is fired and the cat is killed. This whole apparatus is sealed inside a light-proof, radiation-proof box.

The rate of radioactive decay for the large numbers of atoms of the radioactive substance is well known from repeated measurements on prior samples. You, as the expert researcher, are then asked to predict whether, at some given later point in time, the cat in the box is alive or dead.

For Schrodinger the answer was: neither and both. Quantum physics contends that, until the measurement is made (the box is opened and the cat is observed), both conditions exist simultaneously. The cat in the box is both alive and dead, or in some state of suspended animation between the two, until the observation of its condition is made. While this, and other aspects of quantum physics, tend to fly in the face of everyday logic, this story does have significant meaning for the statistical researcher.

As research statisticians we must agree with Schrodinger's interpretation of the events, not because of any quantum measurement that is to be taken, but rather because of the inherently statistical and unpredictable nature of radioactive decay. Prior measurements have established the rate at which *large numbers* of radioactive atoms undergo decay. Such formulae and empirical knowledge are very accurate at determining how much of a given sample will remain radioactive over some period of time. Unfortunately, no statistic known can accurately predict when a single atom will undergo that decay. While that atom will behave within the broad framework drawn for many atoms of its type, we cannot know with any degree of certainty whether it will decay right now or one-hundred million years from now. Given that it will, eventually, decay any prediction of the time of that decay must be random. And, therefore, any attempt to treat that single atom as if it exhibits the properties of a large group of like atoms is inherently flawed.

Extreme Points

The existence of an outlier data point, regardless of how extreme it is, does not necessarily violate the distributional assumptions under which the linear least-squares model operates. The Gaussian distribution allows for such extreme data points, though ascribing them a necessarily low incidence of occurrence. The rationale guiding the elimination (or recoding) of that extreme point should not be one of low probability of occurrence, but rather interference in the stability of the regression model parameters (due to inflated data variability).

In this situation a comparison between a model containing that extreme point, and one with the point eliminated or recoded, can be done. If there is meaningful difference between the interpretations of the two models the researcher must contextually decide if the reduced model is the best interpretation of the data relationships. In that case the extreme point, though viable statistically, hinders understanding by its inclusion in the model.

With more than one potential outlier data point the problem becomes quite complex. The researcher must answer not only the question above, but also whether the location of the outlier points correspond with the anticipated distributional assumptions. Such violations are the basis upon which robust regression techniques are built (Rousseeuw & Leroy, 1987).

When the researcher is confident in their assumptions, however, a more conservative (though not necessarily statistically reliable) strategy is in order. Failing to remove any extreme data points allows their influence to remain, and acknowledges that they are a proportionally representative part of the data set. Eliminating too many extreme points, especially if all accomplished in a single pass, will reduce data set variability and increase predictability with the sacrifice of costly data and important power. It also ignores the reality that, with the Gaussian distribution, extreme data points can (and do) exist.

An alternative, proposed strategy would be to iteratively reduce the data set by the most influential data point, one at a time. This would allow the researcher the opportunity to recalculate the regression and residual statistics without the influence of that data point, perhaps reaching a tenable fit between data and model without too much data set manipulation.

This strategy could be applied to the final data example presented in this paper. Traditional outlier treatment methodologies would argue for the elimination of all six of the identified outlier points in one shot. An iterative technique, however, would begin by removing the most influential data point first, then continue with recalculating parameters and reexamining the model. If this

were done in this example one would fine that the intercept of the least-squares model shifts downward slightly (to accommodate those outlier points below the line). This new model is a much better fit to the data and, although somewhat differently interpreted that the first model, does not suffer from the extreme outlier cases that the first did. [Note: the DrawReg program has the option to iteratively reduce a data set, automatically or through manual entry, until some predetermined criteria are reached.]

Conclusion

An iterative-reduction technique can be criticized for manipulating the data to seek the fit to a particular model. However, all residual analysis and correction techniques do this. It is the very nature of outlier analysis to examine extremely influential data points and to decide if their inclusion in a proposed model skews that model to a meaningful degree. Towards this end I believe that the iterative technique is the most honest of all, since it presents all of the potential alternatives to the researcher for interpretation.

Unfortunately, few statistical software packages support this kind of analysis. Until less computationally-intensive methods can be found such approaches may be theoretically interesting but of little practical significance. In data sets where a limited number of data points are identified as extreme this may prove to be a moot question. Existing analysis methods previously shown can bring these points to the attention of the researcher, who can then take appropriate action.

It is when there are a larger number of points that are extreme to the model that our techniques appear to be lacking. Robust methodologies provide an alternative, but only when the extreme points can be identified as adhering to other distributional assumptions than Gaussian. A test of location and density is needed for these extreme points, much like the non-parametric Kolmogorov-Smirov Z is used for the core of data, to determine if the outlier data is conforming to Gaussian expectations. [Note: DrawReg compares the rate of outliers identified to the rate expected under a normal distribution, though does not computer the density of these points in relation to the expected density.]

The use of such a test could indicate to the researcher whether to approach these points as potentially non-proportionally representative of a normally distributed population or as data representative of some other distributional qualities. It is to the future to create an easily-implemented method for that determination.

References

Draper, N.R. & Smith, H. (1981). *Applied Regression Analysis (2nd Edition).* New York, New York: John Wiley & Sons.

Harnett, D.L. (1982). *Statistical methods (3rd Edition).* Reading, MA: Addison-Wesley.

Hays, W.L. (1988). *Statistics (4th Edition).* Fort Worth, Tx: Holt, Rinehart and Winston.

Hoaglin, D. & Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician, 32,* 17-22.

Kleinbaum D.G., Kupper, L.L., & Muller, K.E. (1988). *Applied regression analysis and other multivariable methods (2nd Edition).* Boston, MA: PWS-Kent.

Microsoft Corporation (1989). *Microsoft QuickBASIC 4.5.* Redmond, WA: Author.

Norusis, M.J. (1990). *SPSS/PC+ Statistics 4.0 for the IBM PC/XT/AT and PS/2.* Chicago, IL: SPSS Inc.

Rae, A. (1986). *Quantum physics: Illusion or reality?* London, England: Cambridge University Press.

Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust regression and outlier detection.* New York, New York: John Wiley & Sons.

Spatz, C. & Johnston, J.O. (1989). *Basic statistics: Tales of distributions (4th Edition).* Pacific Grove, CA: Books/Cole.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix A

The DrawReg program takes data from a user-supplied data file and provides case-by-case data point manipulation in the creation of a linear regression model. General format for running DrawReg is:

DRAWREG DATAFILENAME

Where the DATAFILENAME is the filename of the ASCII (text) data file containing the user's data.

The data points in this file may be either integer or real, and can be input 'free format' with one case to a line. The first value on each line is taken as the X data point, while the second is as the Y point. Up to a maximum of 500 data points may be analyzed, with a minimum of 30 data points necessary for the program to run.

The program has been designed to run on any DOS compatible PC computer system as long as that system is capable of displaying graphics. DrawReg will automatically detect the highest resolution graphic screen present (VGA, EGA, HGC, or CGA) and will adjust the display for the best resolution possible for that screen type. DrawReg has several programmed function keys under user control:

<F1>	Displays the help screen information.
<S><F1>	Toggles the blinking of the outlier data points.
<F2>	Removes the next largest potential outlier case (see <S><F5>).
<S><F2>	Removes a user-entered observation from the data set.
<F3>	Restores the last data point removed back into the data set.
<S><F3>	Restores a user-entered observation back into the data set.
<F4>	Automatically removes data points (<F2>) criteria set (<S><F6>) is met.
<S><F4>	Automatically restores all of the data points in the original data set.
<F5>	Allows changes to the Z-critical value. Touch this key once, then use the <UP-ARROW> and <DOWN-ARROW> keys. Touch <F5> to exit.
<S><F5>	Chooses standardiZed, Studentized, Cooks, or Leverage case selection.
<F6>	Toggles the display of the X-Y means crosshair.
<S><F6>	Chooses criteria of % of z-table or # value below critical.
<F7>	Toggles the display of the current regression line.
<F8>	Toggles the display of the current prediction bands.
<F9>	Displays the numerical statistical values.
<F10>	Displays the current list of potential outliers.
<ESC>	Ends DrawReg.

For the more technically minded, the following formulae are used in the DrawReg program. Citations to where these formulae can be found are given next to each equation:

Mean of X: $\bar{X} = \frac{\Sigma X}{n}$ (S & J, p. 36)

Sdev of X: $s_x = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n - 1}}$ (S & J, p. 60)

Correlation: $r = \frac{\Sigma XY - \frac{\Sigma X \, \Sigma Y}{n}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{n})(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n})}}$ (K, K, & M, p. 80)

Regression Y on X: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (K, K, & M, p. 51)

where: $\hat{\beta}_1 = r\,(\frac{s_Y}{s_X})$ (K, K, & M, p. 80)

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ (K, K, & M, p. 51)

Serr of Y|X: $s_{Y|X} = \sqrt{(\frac{n - 1}{n - 2})(s_Y^2 - \hat{\beta}_1^2 s_X^2)}$ (K, K, & M, p. 56)

Y predicted: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (H, p. 477)

Y residual: $Y_{Res} = Y_i - \hat{Y}_i$ (H, p. 496)

Prediction Bands: $\hat{y}_i \pm z_{\alpha/2}\, s_{\eta X} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)\, s_X^2}}$ (K, K, & M, p. 62)

Standardized residual: $Z_{Y_{Res}} = \frac{Y_{Res}}{s_{Res}}$ (K, K, & M, p. 186)

where the critical z is: User selected (follows Z dist for n>50)

Leverage: $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)\, S_X^2}$ (K, K, & M, p. 199)

where the critical h is: $h_i > \frac{2\,(k+1)}{n}$ (H & W, 1978)

[Note: Mahalanobois' Distance (D^2) was not used since D^2 is proportional to $(n-1)h_i$ when n>50. Also, note that SPSS/PC+ calculations of leverage will differ from DrawReg's since SPSS/PC+ uses a "centered" leverage (Norusis, 1990, p. B-106).]

Studentized residual: $r_i = \frac{Y_i - \hat{Y}_i}{s\sqrt{1 - h_i}}$ (K, K, & M, p. 186)

where the critical r is: User selected (follows Z dist for n>50)

Cook's Distance: $d_i = \frac{(Y_i - \hat{Y}_i)^2\, h_i}{(k+1)\, s^2\, (1 - h_i)^2}$ (K, K, & M, p. 201)

where the critical d is: $d_i > 1$ (S, p. 94)

Z table (Standard normal curve probabilities) (K, K, & M, p. 644-5)